Understanding Matrix Assisted Continuous Cocrystallisation using Data Mining approach in Quality by Design (QbD)

Billy Chabalenge[2], *Sachin Korde*[1,2], *Adrian L Kelly*[1,3], *Daniel Neagu*[3], *Anant Paradkar*[1,2*]

AUTHOR ADDRESS

1. Centre for Pharmaceutical Engineering Science, University of Bradford, Bradford, UK.

2. School of Pharmacy and Medical Sciences, University of Bradford, Bradford, UK

3. Faculty of Engineering & Informatics, University of Bradford, UK

ABSTRACT

The present study demonstrates the application of decision tree algorithms to the cocrystallization process. 54 batches of Carbamazepine–Salicylic acid cocrystals embedded in poly(ethylene oxide) were manufactured by hot melt extrusion and characterised by PXRD, DSC and NIR. This data set was then applied in WEKA, an open-sourced machine learning software to study the effect of processing temperature, screw speed, screw configuration and poly(ethylene oxide) concentration on the percentage of cocrystal conversion. The decision trees obtained provided statistically meaningful and easy to interpret rules, demonstrating the potential to use the method to make rational decisions during cocrystallization process development.

INTRODUCTION

Cocrystallization is a novel approach used to design and develop new multi-component crystalline solids with unique bulk properties. Engineering of poorly soluble active pharmaceutical ingredients (APIs) by cocrystallization leads to superior solubility, stability, better release and thus better bioavailability without modifying the API's physiological action

[1, 2]. Studies have also shown that cocrystals can be helpful in dosage form formulation as the multicomponent structure of cocrystals can offer improves mechanical properties, give rise to better powder flowability and compressibility [3, 4].

Traditional methods for cocrystallization which are extensively used in screening novel pharmaceutical cocrystals include mechanical grinding, cooling and antisolvent addition methods, slurry conversion and solvent evaporation [5, 6]. These methods are however difficult to scale up and apply commercially due to the requirement of high energy input (during grinding) and the need to control organic solvents and crystallization conditions such as supersaturation and concentration of compounds[7]. On the other hand, cocrystallization by hot-melt extrusion (figure 1) has been shown to be advantageous as a fast, continuous, solvent-free and readily scalable method for cocrystal synthesis[8]. Furthermore, the number of production steps to produce the final dosage form can be reduced through hot melt extrusion.

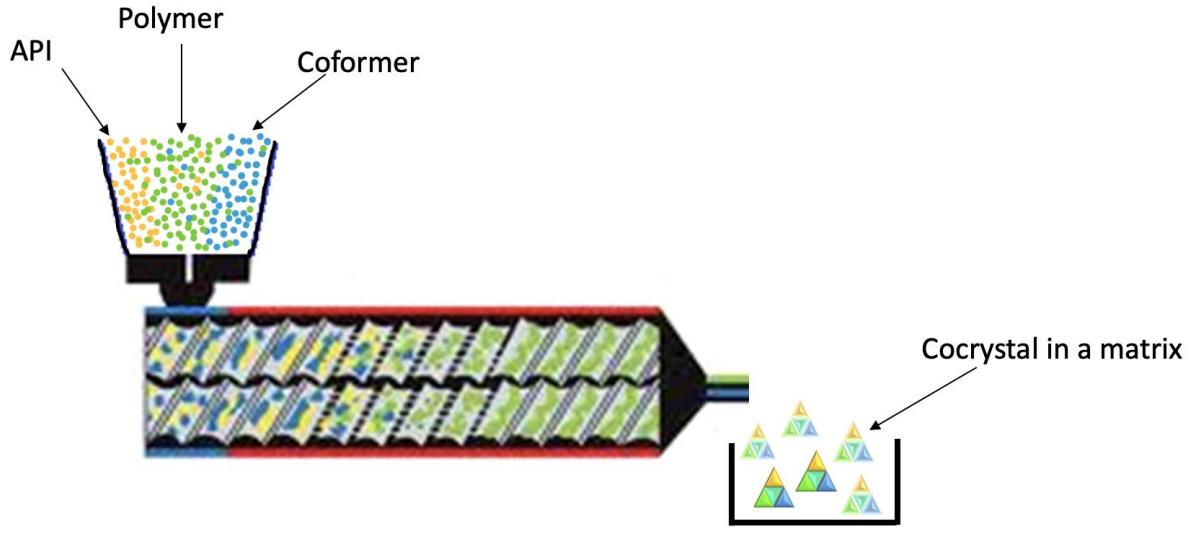


Figure 1. Schematic diagram of hot melt extrusion.

A matrix or polymer can be added in small quantity by weight fraction to the mixture of the API and coformer during hot melt cocrystallization. This process is called matrix assisted cocrystallization (MAC) as the manufactured cocrystals are embedded in the polymer [9].

During extrusion, the matrix acts analogous to a catalytic solvent because the molten/softened polymer induces intimate mixing of the API and coformer and also reduces excess shear stress which occur as a result of extruding dry solids. In the final MAC product, the polymer is a functional component which influences the properties of the obtained MAC product such as improved compressibility and flowability. MAC was first demonstrated in 2014 by Boksa et al9 who simultaneously produced and formulated Carbamazepine – nicotinamide cocrystals in Soluplus® as a matrix. To date, limited number of API – coformer in a polymer studies by hot melt extrusion cocrystalisation process exist in literature. The polymers used so far in available literature on hot met extrusion based cocrystalisation include; Soluplus®, Eudragit EPO, Vinyl acetate copolymer (PVPVA64), poloxamer P407, hydroxypropylmehtylcellulose acetate succinate (HPMCAS), Kollicoat IR® and poly(ethylene oxide) (PEO) 7, 9, 10. The use of polymers as matrix in other solid dispersion techniques has also been well documented 11-13.

Despite hot-melt extrusion being a continuous process and allowing several downstream options to produce various pharmaceutical preparations, there is still limited understanding of the critical processing and formulation parameters as evident by the limited number of formulations receiving regulatory approval [14]. The introduction of quality by design (QbD) in the pharmaceutical industry has helped to understand both formulation and process-related attributes during product development. In manufacturing of new or marketed products, QbD can help in pre-determining the risk potential of various operations, assuring that suitable control strategies can be applied on time. Since QbD is a science-based approach, it provides a basis for optimizing and improving the manufacturing operation by employing design of experiments (DoE) to the drug, co-former and excipient attributes and process parameters in order to determine the critical quality attributes and define a design space. The response surface methodology technique is used in DoE to optimise the process by producing a

polynomial equation describing the relationship between inputs and outputs. The advantage of this model is that it is easily understood; however, relationships between inputs and outputs are usually complex and can result in poor estimation of the processing conditions.

Recent advancements in computer science and mathematics have helped in developing software that aid analysis of complex data throughout DoE, optimization and/or creating rules within a design space. One of the best known open source software for data mining is called the Waikato Environment for Knowledge Analysis (WEKA[15]) developed by the University of Waikato in New Zealand. It contains a number of well-known algorithms for pre-processing data, classifying, clustering, finding associations in data and also tools for visualizing variables[16].

Nowadays there are many software packages that offer collections of machine learning algorithms for data analysis and data mining. Among these options, the authors preferred the WEKA package as an academic, open source, freely available comprehensive collection of techniques for data pre-processing and modelling. It provides through a user-friendly interface access to the most known machine learning algorithms. As a Java, open-source package under GNU General Public License, it also offers the opportunity to develop further a personalised user interface, should the authors consider to develop and deploy it as a follow-up project, while maintaining control over the data. Other more specialised environments were considered initially, e.g. KNIME, RapidMiner, Alteryx, SAS Miner, Oracle ODM, but most are complex high-level workflow environments that require additional efforts for implementation and maintenance that would have slowed down the current research. Among the algorithms in WEKA, decision trees such as the J48 tree, Reduced error pruning tree (REPtree) and Random trees can be useful in pharmaceutical development as they aid understanding of the cause-effect relationship among attributes by generating easily understood rules that can be used to optimize the process. The ability of decision trees to

convert large complex process data sets into information-rich rules and simultaneously their ability to identify the most significant factors affecting the product quality[17] make them suitable for quality improvement of existing pharmaceutical products and processes, according to QbD principles[18]. In cocrystallization, decision tree rules can therefore be used to select a set of parameters or help in properly adjusting controllable variables during production with the aim of yielding high quality cocrystals. This means that a decision tree predictive model would reduce cocrystallization process development time by helping select correct operation conditions required to yield high quality co-crystals.

The J48 decision tree measures the randomness of information (information entropy) in a data set to determine which attributes are most predictable. This attribute becomes the root of the decision tree and it is split into nodes (rules for an attribute) until there is no further information gain. The five algorithmic steps involved in splitting an attribute are summarised in a flowchart in figure 2:

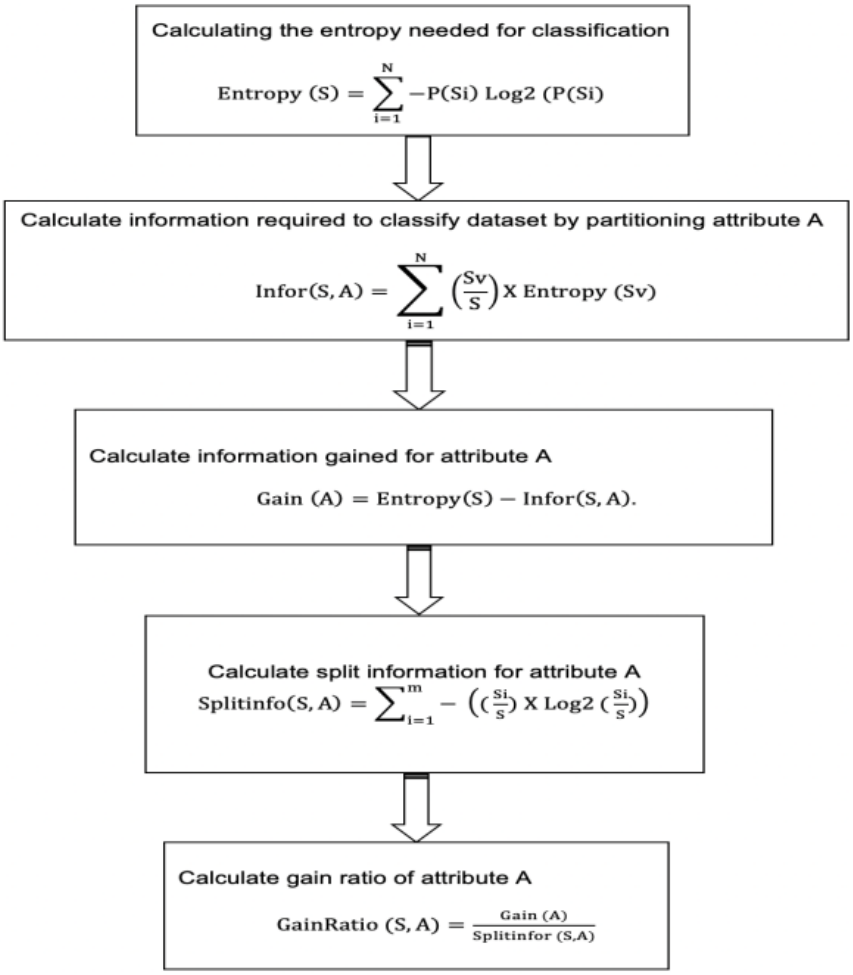


Figure 2. Major steps in calculating information entropy

Through these steps, a decision tree is built by the J48 algorithm by using the "divide and conquer" approach that is being greedy towards a variable that maximise GainRatio during splitting of data sets. The advantage of the J48 is that it has a transparent mechanism and easy to interpret solutions from the tree[19].

The REPtree algorithm works like the J48 tree to build either a decision or regression tree from the information in a data set, then prunes the tree by reducing the error. Pruning starts with replacing the leaf nodes with the most frequent class to the root node and if accuracy is not affected, the replacement is kept. Pruning in this manner continues and only stops when accuracy is decreased. Even though the REPtree algorithm is somewhat naïve, it has an

advantage of being simple to understand and faster at producing a regression or decision tree [19, 20].

Random tree algorithms are supervised form of decision trees that use the idea of bagging (combines weak attributes to build one strong decision tree) to build a decision tree which is different from standard decision tree classifiers. In building a regression or decision tree, the Random tree first samples the training dataset through bagging to produce individual random decision trees (small trees). Each of the randomly created trees contains attributes generated from different subsets and samples. To compute the final decision, each tree has a "vote" and the most frequent variables are optimised through a random procedure to produce the overall classification [21, 22].

Pharmaceutical literature from the past few years shows that analysing manufacturing processing data in order to gain knowledge required for improving product quality is becoming increasingly important especially with the emphasis of building quality into a product rather than testing it[23, 24]. Data models have further been shown to be helpful in simulating pharmaceutical processes such as wet and dry granulation[25, 26] and helpful in selecting appropriate excipients and manufacturing process condition and further, in optimising critical processing parameters. Therefore, application of algorithms in knowledge gaining for manufacturing processes such as hot melt extrusion is justifiable. Further, there is no literature that shows the application of decision tree models in understanding effects of processing variables and material attributes on cocrystallization process. The present study was focused on showing the application of decision tree algorithms in WEKA to build models that can be used to describe hidden relationships in cocrystallization data sets in order to optimize cocrystal conversion rate. WEKA has a comprehensive collection of techniques for data pre-processing and modelling, to the best of our knowledge, none has ever tried to use WEKA as a data mining tool on cocrystalisation data set.

- **EXPERIMENTAL SECTION**

Materials.

Carbamazepine (CBZ) was purchased from Taj pharmaceuticals, India (Lot no. TPL/CARB/002), salicylic acid (SAL) was purchase from Sigma-Aldrich (UK), polyethylene oxide (PEO N80 grade) was purchased from Colorcon (UK). All the organic solvents used for HPLC analysis were purchased from Sigma-Aldrich (UK).

Variables.

A sizable amount of data set is required during data mining as data is split into training and testing dataset. Therefore, 54 experiments were designed as data sets using the custom-design option in Design-Expert® v11.0.5. software considering screw speed, barrel temperature, matrix concentration and screw configuration as variables. The input variables used to conduct the 54 experiments are shown in Table 1. Batches of 1:1 molar ratio mixture of Carbamazepine and Salicylic acid were accurately weighed, mixed and 5%, 10%, and 15% PEO N80 concentrations were added to different batches and then mixed for 10 minutes in a turbula mixer. The homogenous mixture was then fed manually and extruded using a co-rotating twin screw extruder (Pharmalab 16, Thermo Fisher Scientific) with a length-to-diameter ratio of 40:1 without a die using two different screw configurations (SI). During the extrusion process, temperatures in zones 6 to 10, screw speed and screw configuration were changed after collecting about 50g of each batch. SI 2 shows the temperatures in other zones, which were kept constant for all experiments.

Table 1. Input variables used for the custom-design of experiment

Variable	**Level (- 1)**	**Level (0)**	**Level (+)**
Screw speed (rpm)	50	75	100
Barrel temperature (^{o}C)	110	115	120

Polymer concentration (%)	5	10	15
Screw configuration	Configuration 1*	-	Configuration 2*

*Provided in SI 1

Powder X-ray diffraction.

The obtained matrix assisted cocrystals were characterised by X-ray diffraction, differential scanning calorimetry and near infrared spectroscopy. X-ray powder diffraction (PXRD) was performed using a Bruker D8 diffractometer, operating at a voltage of 40kV Cu Source, amperage of 40mA and wavelength of 1.54Å to assess crystallinity of the obtained cocrystals in a matrix. All samples were first ground using a motor and pestle and then loaded evenly onto either aluminium or silica sample holders. Scanning was then done at a 2θ of 5^o – 30^o using a 0.02° step width and 1s time count. The data collected was processed using PowDLL Converter v2.82.0.0 and Origin Pro8 SRO v8.0724 software.

Differential scanning calorimetry (DSC)

Thermal analysis of all samples was generated in the temperature range of 20 – 200^{o}C using a TA instruments Q2000 differential scanning calorimeter with RCS90 cooling unit. About 1.5 – 2.5mg of each sample was weighed accurately and placed in aluminium pans. An empty aluminium pan and lid was weighed accurately and used as a reference. All thermograms were recorded at a heating rate of 10°C min^{-1} in an inert atmosphere by purging nitrogen gas at a flow rate of 50 ml/min. The data collected was processed using MS excel 2016 and TA Universal Analysis 2000 V4.5A software.

Near Infrared Spectroscopy (NIR)

Off-line NIR spectra were taken using an FT-NIR analyzer (Antaris II, Thermo Scientific, UK). Each sample was placed in a transparent glass vial then placed on the NIR integrating sphere and scanned every 30 seconds in the range of 4500 – 10000 cm^{-1} wavenumbers averaging 32 spectra and at a resolution of $8cm^{-1}$. The obtained spectra were analysed using Spectragryph v1.2.12 and MS excel 2016.

Data mining procedure

In this project, three decision tree algorithms were applied to explore the impact of processing parameters and matrix effect on cocrystal conversion. The results obtained from PXRD, DSC and NIR analysis were separately analysed, converted to percentages of cocrystal conversion and used as output data sets. This was done as follows; for PXRD, the ratio of the 2θ peak characteristic of the cocrystal to the ratio of the 2θ characteristic of carbamazepine was used to calculate the percentage cocrystal conversion. For DSC, the enthalpy of the manufactured cocrystals in the matrix and of the pure cocrystal were normalised and used to calculate the percentage enthalpy. For the NIR data, the wavenumber where the carbamazepine peak was at zero, while that of the cocrystal was high and the wavenumber were the carbamazepine peak was high whilst the cocrystal peak was at zero was used to calculate the conversion ratio.

The J48, REPtree and Random tree algorithms in WEKA software v3.8.3 were used for developing models. As these algorithms cannot handle numerical output, the PXRD, DSC and NIR data sets were classified into three categories using the same scale shown in Table 2.

Table Error! No text of specified style in document.**.** Categories used to classify output data for the algorithms

Category	Applied to
High	% cocrystal conversion between 70 – 100%

Medium	% cocrystal conversion between 50 – 69%
Low	% cocrystal conversion between 0 – 49%

The data sets obtained (SI3) were converted to the arff file format suitable for WEKA and loaded into WEKA engine. Pre-processing of data was then performed using the synthetic minority oversampling technique (SMOTE) in WEKA to balance all the three datasets (PXRD, DSC and NIR). Imbalance in dataset may cause model overfit and poor performance of the algorithms[27, 28]. It was therefore important to apply this technique during the pre-processing stage. The SMOTE filter helps in reducing model overfit by balancing classes through creation of synthetic instances on the minority class hence the distribution is balanced. This filter starts by selecting an instance from the minority class randomly and then finds its nearest K-minority class neighbour to produce a synthetic class randomly from a combination of the minority class selected randomly and its neighbour [27]. For PXRD data, the SMOTE filter was applied twice (at 100% and 50%), for DSC data four times (3 times at 100% and once at 65%) while for NIR it was applied twice (at 100% and 40%). The other configurations in the filter were kept at default settings. The three algorithm were then applied on each data set using the default settings of the 10-fold cross validation as a standard evaluation technique. Finally, the algorithms were evaluated for their performance to select the best model rules for each dataset. This was done by comparing their accuracy, precision, recall, F-measure, ROC area and the confusion matrices. Detailed descriptions of these performance evaluation techniques can be found in literature[29-31].

- **RESULT AND DISCUSSION**

54 different experiments were performed based on design expert to study how cocrystal conversion by percentage was affected by different polymer concentrations, screw configurations, changes in barrel temperature and screw speeds. The selection of PEO N80 as a polymer was based on it being semi-crystalline synthetic polyether and having a low grass transmission temperature (- 67oC) convenient for processing during extrusion 32. Also, the functional group in PEO N80 does not lead to formation of amorphous mixture of the starting materials as there is no interaction between the matrix and component of the cocrystal 10. PEO N80 was added in the beginning of the process to facilitate intimate mixing of the drug and coformer. Preliminary studies and literature provided the basis for setting the levels of PEO at a concentration of 5 – 15% 10. Adding a polymer in small quantities based on weight fraction in the mixture helps abate the shear stress during hot met extrusion process thereby reducing the damage to the formed cocrystals 5. Approach was use of PEO N80 as melt processing solvent where during extrusion process due to low glass transition temperature carbamazepine-salicylic acid formed low eutectic system in molten PEO N80 and help nucleate cocrystal seed at that temperature. The 54 extrudates were characterised by PXRD, DSC and NIR.

The PXRD patterns for the starting materials and the cocrystal obtained are provided in SI4. All 54 batches produced 2θ peaks at 6.48, 8.84, 13.08 and 16.44° characteristic of Carbamazepine/Salicylic acid cocrystal as reported in literature10. A 2θ peak at 13.5 (figure 3) is characteristic of pure carbamazepine and this peak was used as an indicator of the residual pure carbamazepine in the formed cocrystals. The overall product obtained is a complex mixture containing some amorphous components and unreacted carbamazepine as well as co-former and traces of impurities such as iminostilbene which makes its analysis complex. Therefore, in the literature the peak intensity ratio has been used in many instances especially when it is for comparative purposes, we adopted the same method. The cocrystal

percentage conversion was calculated from PXRD data by taking the ratio of the area at 2θ peaks 6.48 (cocrystal) to 13.5 (Carbamazepine) as previously used by Ibrahim et al.33 and Kelly et al.34 to determine the relative conversion of the pure drug into cocrystals.

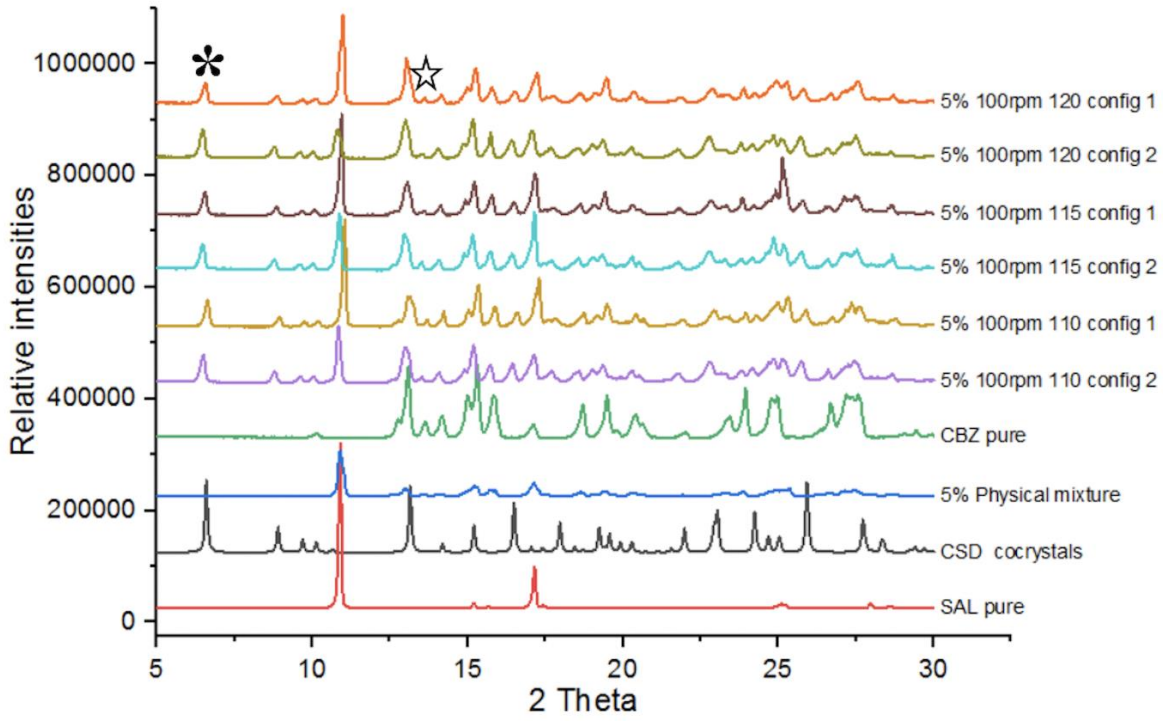


Figure 3. Representative PXRD patterns of the obtained from the 5% MAC products. (*) and (☆) in part C indicates peak characteristic of Cocrystals and Carbamazepine respectively.

The percentage cocrystal conversion is shown in figure 4. The percentage conversion of the pure drugs to cocrystals ranged between 14% to 84%. It can be observed that the percentage conversion increases from the cocrystal containing a low concertation of the polymer (5%) to the ones with higher polymer concentration (15%).

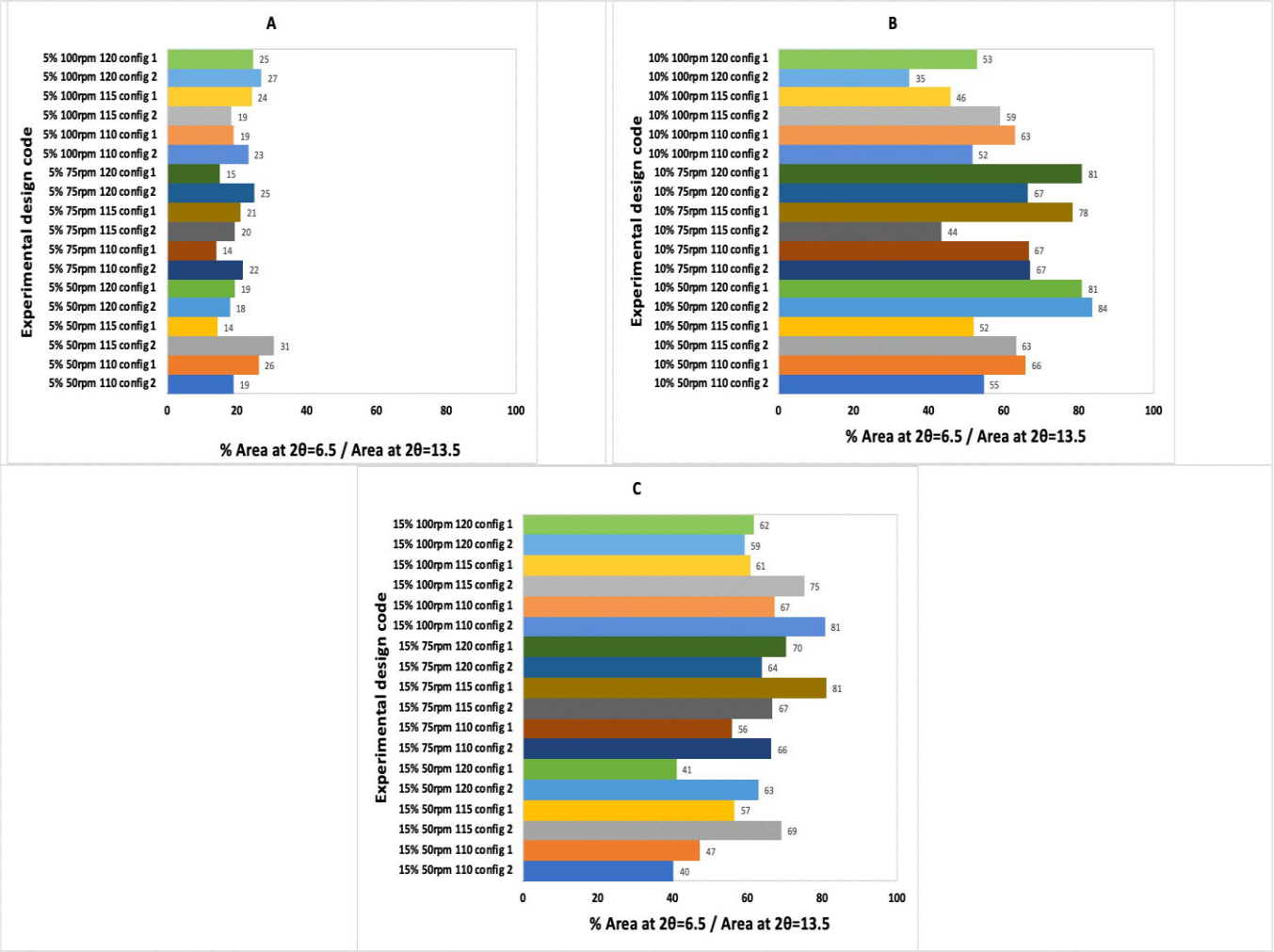


Figure 4. MAC percentage conversion from ratio of PXRD cocrystal peak to the carbamazepine residuals for all processing parameters. A represents all MAC products containing 5% PEO, B for MAC products containing 10% PEO and C for 15% PEO batches.

The starting materials, one pure cocrystal and produced MAC products were also characterised by DSC (SI5). The enthalpy of the melting peak of the pure cocrystal and MAC were normalised and the ratio of the normalised enthalpy of the MAC product to that of the pure cocrystal was converted to percentage as an indicator of the percentage of cocrystal conversion in the MAC products. The percentage conversion ranged between 23.92% to 80.52% as shown in figure 5.

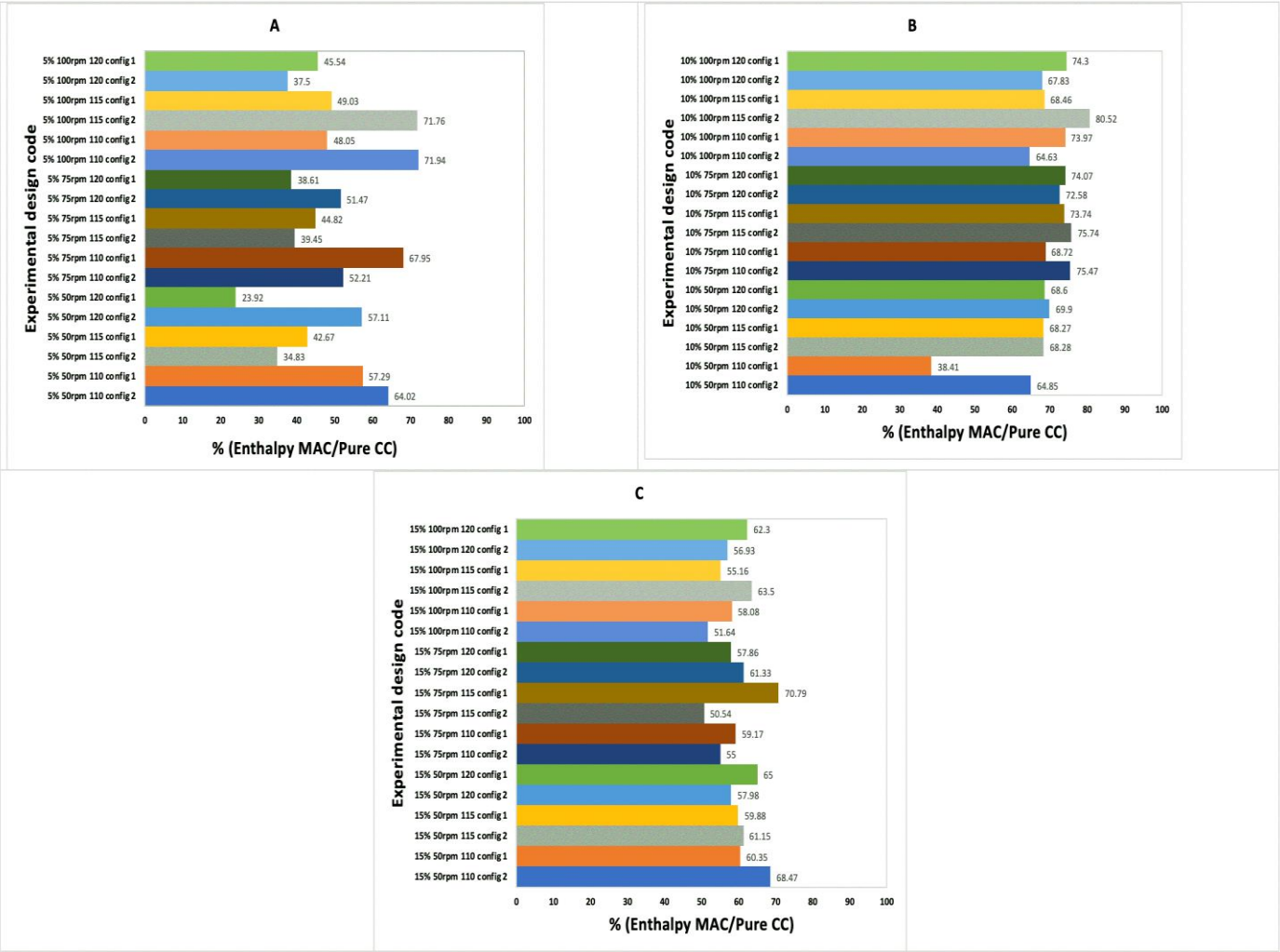


Figure 5. DSC enthalpy percentage for all MAC products from ratio of MAC product melting endotherm to that of the pure cocrystal for all processing parameters. A represents all MAC products containing 5% PEO, B for MAC products containing 10% PEO and C for 15% PEO batches.

These results showed that the MAC containing 10% and 15% PEO generally led to higher cocrystal conversion. The Second Derivative NIR spectral region, 6788.31 – 6958.0183cm^{-1} (SI6) was also used to estimate percentage cocrystal conversion by converting ratio of percentage transmittance of the MAC products to that of the pure carbamazepine. This region was chosen as a good indicator of cocrystal conversion because at wavenumber 6815.31cm^{-1}, pure carbamazepine and MAC extrudates had high percentage transmittance while the pure cocrystal transmittance was zero indicating the presence of residual carbamazepine in the extrudates. Further, at wavenumber 6907.877cm^{-1} the carbamazepine percentage

transmittance was at zero while that of the pure cocrystal and extrudates was high as presented in figure 6 on products containing 5% PEO which were extruded at 100rpm at the three different temperatures.

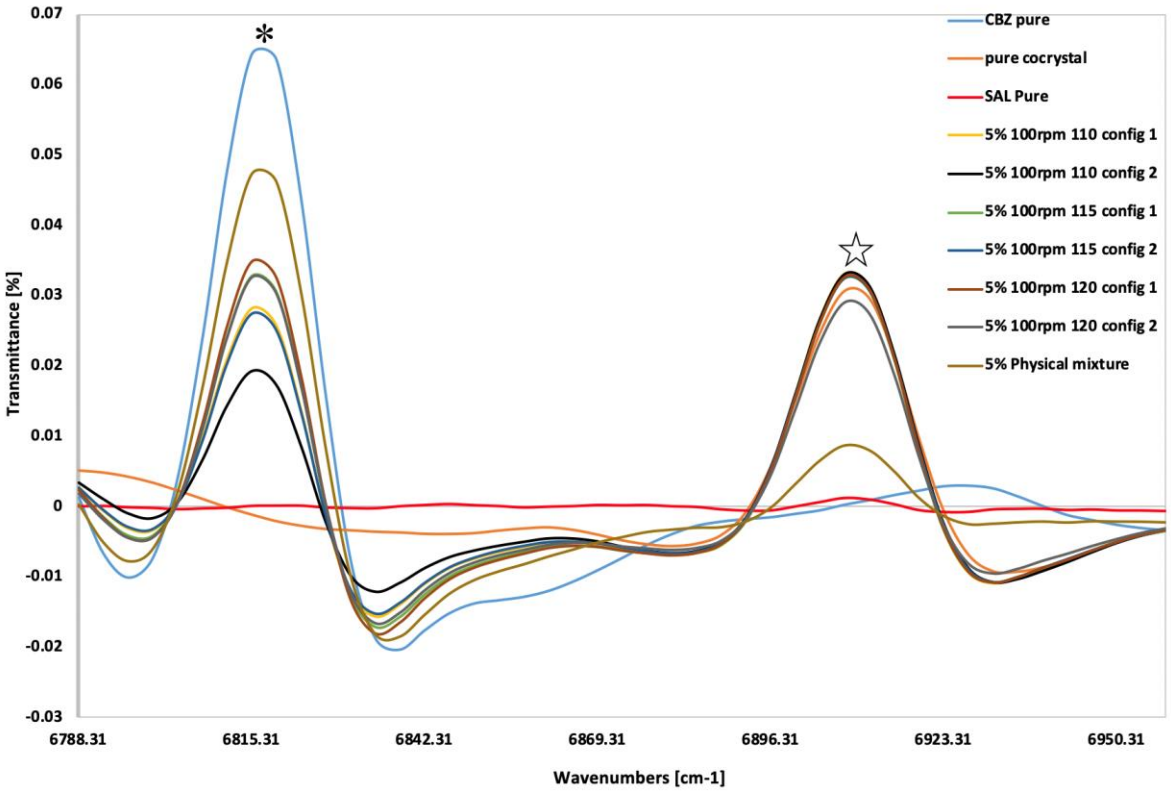


Figure 6. Representative of second derivative NIR spectra obtained for the 5% MAC products. (*) represents the wavenumber where the percentage transmittance of pure carbamazepine and extruded products were high and (☆) represents the wavenumber where the percentage transmittance for the pure cocrystal and extrudates were high.

The percentage conversion results of all 54 batches are shown in figure 7. Similar to the PXRD and DSC results, the products containing 5% of the polymer showed lower percentages of cocrystal conversion.

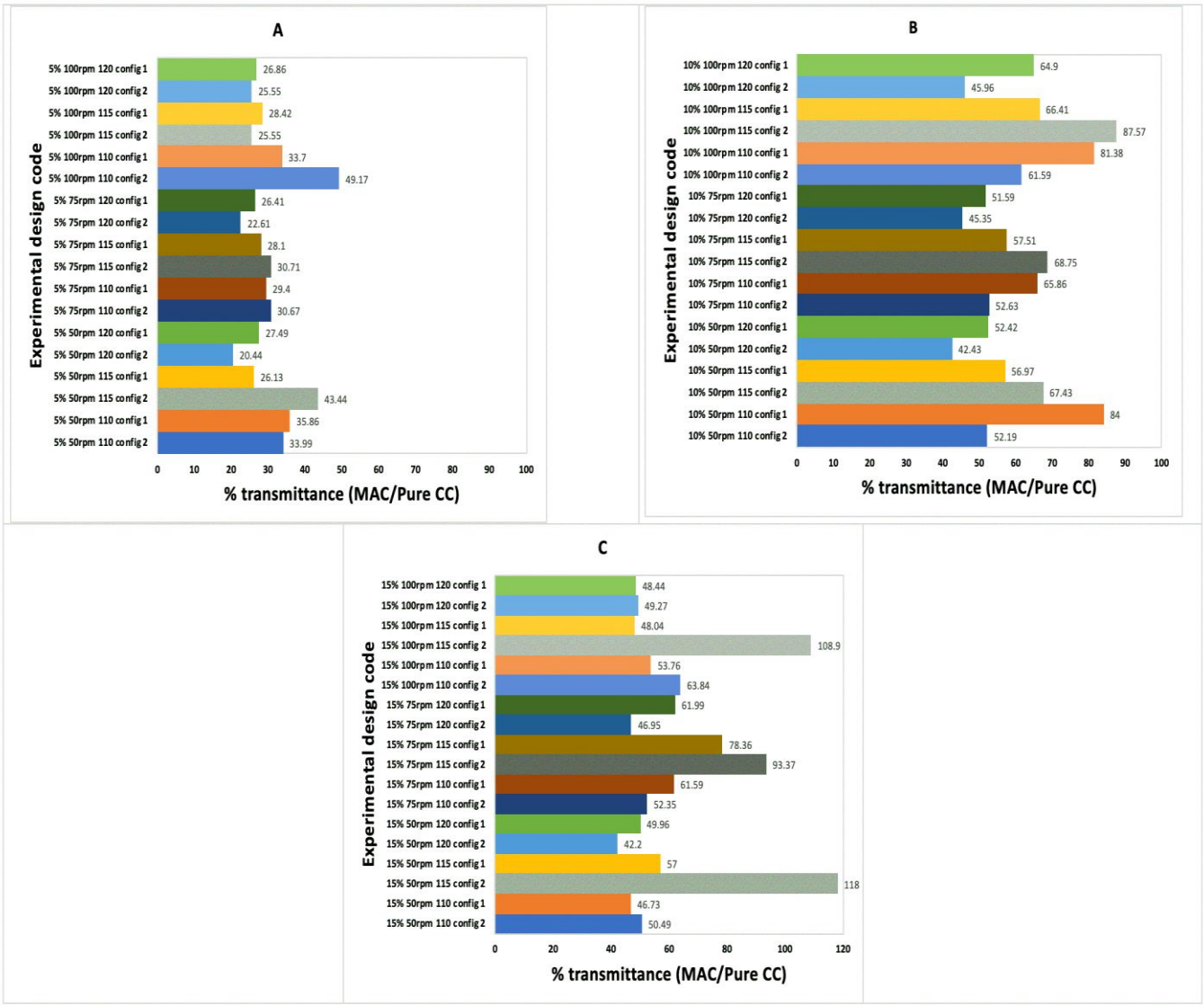


Figure 7. MAC percentage conversion from ratio of NIR cocrystal peak to the carbamazepine residuals for all processing parameters. A represents all MAC products containing 5% PEO, B for MAC products containing 10% PEO and C for 15% PEO batches.

To predict cocrystal conversion rate in a given set of material and processing parameters, three decision trees in WEKA software (J48, REPtree and Random tree) were used for data mining on each data set. The statistical performance results of the three decision trees on the datasets are shown in table 3. From these results, the J48 had higher percentages in terms of correctly classified instances (74.29%, 81.13% and 81.82% on PXRD, DSC and NIR data respectively), and also had the lowest percentage of incorrectly classified instances on all the

three data sets. The other statistical parameters provided in table 3 varied between the three algorithms on each data set. The J48 decision trees were therefore were chosen to visual the influence of processing parameters and polymer on cocrystal conversion rate for each data set.

Table 3. Statistical performance of decision trees on PXRD, DSC and NIR data.

	Algorithm	CCI	ICI	TP	FP	Precision	Recall	F-Measure	ROC area
XRD data	J48	74.29%	25.71%	0.750	0.065	0.857	0.750	0.800	0.845
	REPtree	70.00%	30.00%	0.750	0.022	0.947	0.750	0.837	0.827
	Random tree	70.00%	30.00%	0.792	0.087	0.826	0.792	0.809	0.852
DSC data	J48	81.13%	18.87%	0.861	0.114	0.795	0.861	0.827	0.900
	REPtree	80.19%	19.81%	0.889	0.157	0.744	0.889	0.810	0.883
	Random tree	72.64%	27.36%	0.694	0.029	0.926	0.694	0.794	0.833
NIR data	J48	81.82%	18.18%	0.815	0.051	0.917	0.815	0.863	0.900
	REPtree	66.67%	33.33%	0.889	0.128	0.828	0.889	0.857	0.910
	Random tree	71.21%	28.79%	0.815	0.103	0.846	0.815	0.830	0.856

Key: CCI = Correctly classified instances
ICI = Incorrectly classified instances
TP = True positive rate
FP = False positive rate

From PXRD data, the influence of the polymer and processing variables on cocrystal conversion are shown in figure 8 for the J8 decision tree. The tree graph model gives insight into the hidden relationship in the dataset which would lead to low, medium or high cocrystal conversion. A polymer concentration of less than 5% should be avoided during cocrystallization as it leads to low cocrystal conversion. The tree gives two best rules to yield high cocrystal conversion in the matrix, however, the best rule shows that using a higher than 5% polymer concentration, screw configuration 1, a barrel temperature of greater than 110oC and a screw speed between 50rpm and 99.88rpm is desirable as 18 instances followed this rule and none was misclassified. The processing parameters screw speed and temperature have multiple and confounding effects on both the dynamics of the process and the

cocrystallisation mechanism. They affect residence time, shear intensity, effect on solubilities of the different components, chances of degradation glass transition temperature of the system and interaction variables too, which makes the process complex and necessary to use data mining tools. Previous studies on extrusion cocrystallisation have found that low screw speeds produced the highest co-crystal yield due to the extended time the materials were exposed to the process conditions 8.

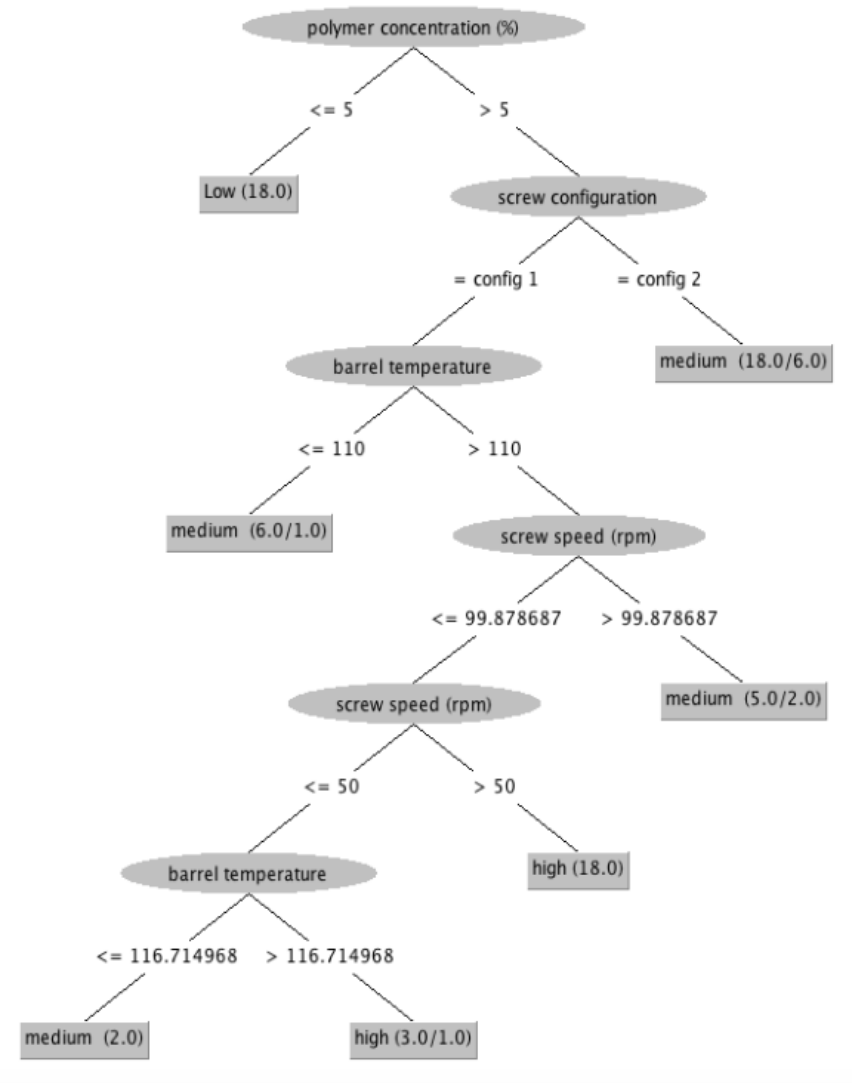


Figure 8. J48 decision tree model generated on PXRD data set.

On NIR data, the tree model produced (figure 9) also gives insight in the hidden relationship between variables. In this model, screw configuration also has an influence on the cocrystal

conversion unlike in the PXRD data and can lead to low – medium cocrystal conversion if correct processing limits are not used for the variables. To yield high cocrystals in a polymer, the best rule given should involve using a polymer concentration of greater than 10% and barrel temperatures of between 110oC and 115oC as this rule was followed by 18 instances.

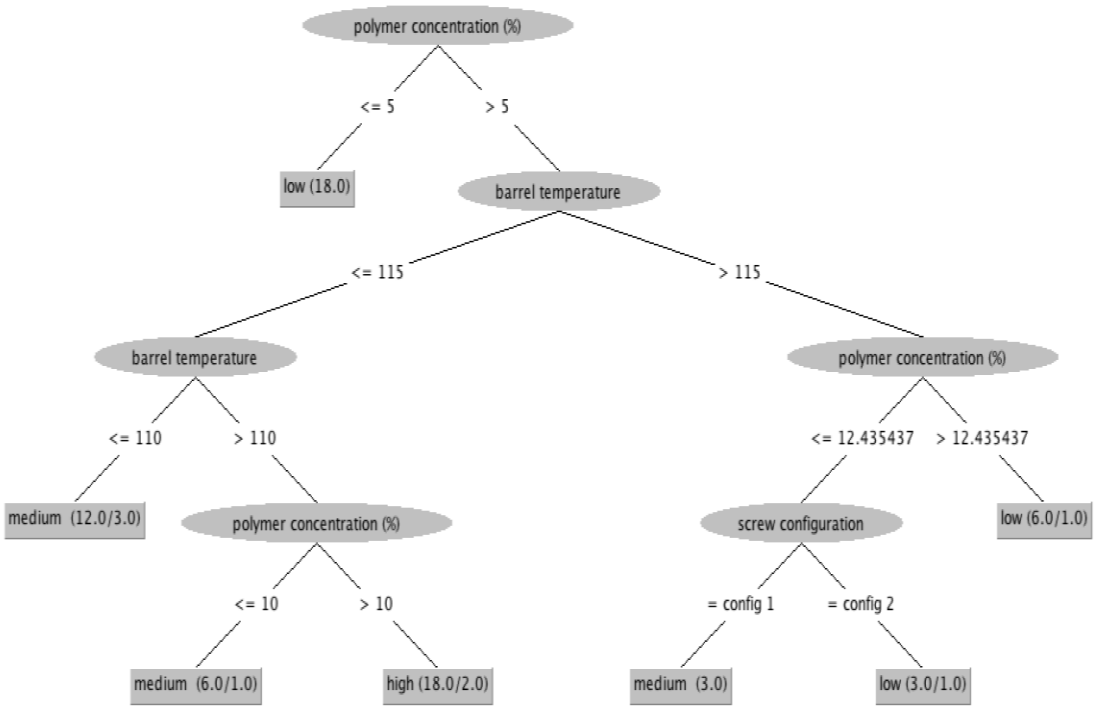


Figure 9. J48 decision tree model generated on NIR data set.

The J48 model produced on DSC data is shown in figure 10. To yield a high percentage cocrystal conversion, a polymer concentration of between 9.05% to 13.97% and screw speed greater than 61.52rpm should be used. Further, a combination of a polymer concentration of less than 9.05%, screw configuration 2 and a screw speed of greater than 83.55rpm can lead to high cocrystal conversion.

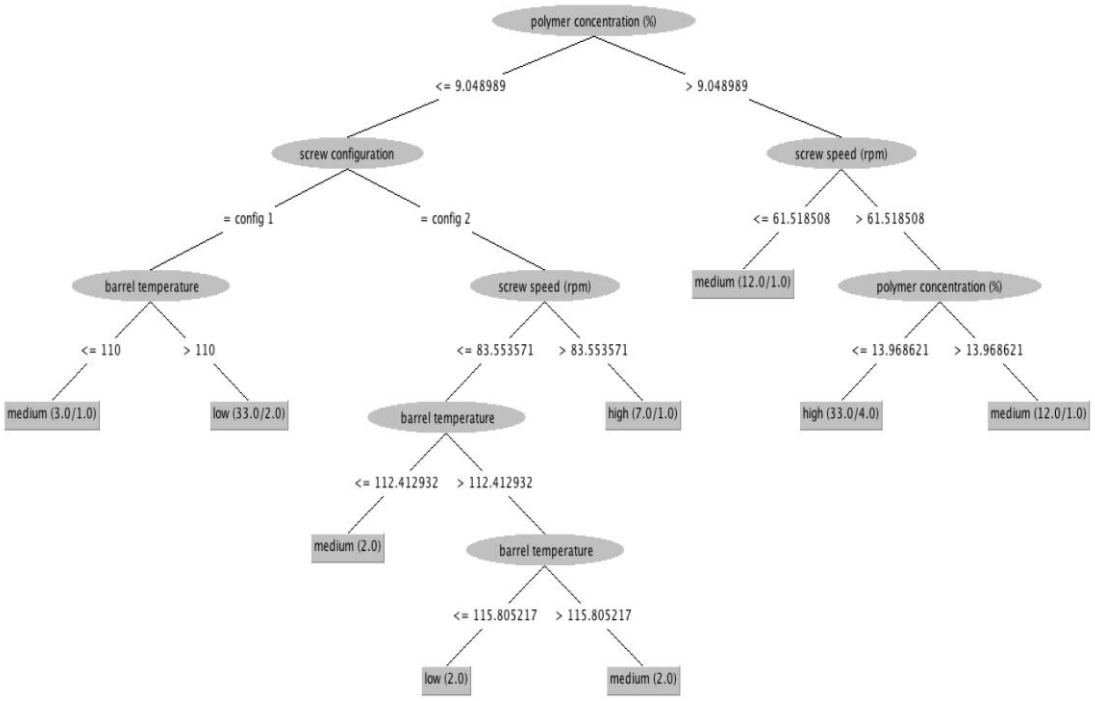


Figure 10. J48 decision tree model generated on DSC data set.

The results of the three decision tree models show that polymer concentration had the highest entropy as it was the root node of the trees. This indicates clearly that cocrystal conversion percentage in a matrix is mainly influenced by the polymer concentration; however, cocrystallization process parameters also affect the yield percentage. These observations are in agreement with other published literature involving the matrix cocrystallization process [5, 7, 9], which confirms how useful decision trees could be in constructing conclusions on the effects of process parameters and polymer on cocrystallization process. Further, these models offer advantages of being easy to visualize the complex cocrystallization process and using the rules in the tree, give a deeper understanding and knowledge of attribute combinations that would yield high cocrystal conversion. Therefore, the decision tree models can be successfully be used as a decision-support tool during process optimisation to achieve high yield of cocrystals embedded in a polymer matrix, subject to consistent validation and assessment.

CONCLUSION

In this study, decision trees were successfully used as a data mining technique to demonstrate knowledge gained on the cause-effect relationship of process parameters and matrix with the aim of yielding high percentage cocrystals embedded in a polymeric matrix. Gaining such knowledge has been demonstrated that decision trees can be a handful tool in making rational decisions from the simple and informative rules obtained during crystallisation process. Further, decision tree approach as a data mining tool can fit within the requirements of quality by design. In future works, there is need of increasing the quantity of data and case studies for comparing decision trees and association mining in WEKA for further knowledge gaining. A comprehensive validation study involving different cocrystal forming pairs, polymers and process conditions and even HME designs would provide interesting outcome by applying data mining techniques.

SUPPORTING INFORMATION

The supporting information is available free of charge at http://pubs.acs.org.

Details of the processing variables and results of characterization of all 54 batches are provided as supporting information.

ACKNOWLEDGEMENTS

Authors would like to acknowledge funding support from Commonwealth Scholarship Commission in the UK (ZMCS-2018-783) and Engineering and Physical Sciences Research Council (EPSRC EP/J003360/1 and EP/L027011/1)

AUTHOR INFORMATION

***Corresponding Author**

Professor Anant Paradkar

Centre for Pharmaceutical Engineering Science, School of Pharmacy, University of Bradford, Bradford, BD7 1DP, UK.

E-mail: A.Paradkar1@Bradford.ac.uk

REFERENCES

(1) Moradiya, H. G.; Islam, M. T.; Halsey, S.; Maniruzzaman, M.; Chowdhry, B. Z.; Snowden, M. J.; Douroumis, D., Continuous cocrystallisation of carbamazepine and trans-cinnamic acid via melt extrusion processing. *CrystEngComm* **2014,** 16, 3573-3583.
(2) Cherukuvada, S.; Kaur, R.; Guru Row, T. N., Co-crystallization and small molecule crystal form diversity: from pharmaceutical to materials applications. *CrystEngComm* **2016,** 18, 8528-8555.
(3) Chattoraj, S.; Shi, L.; Sun, C. C., Understanding the relationship between crystal structure, plasticity and compaction behaviour of theophylline, methyl gallate, and their 1 : 1 co-crystal. *CrystEngComm* **2010,** 12, 2466.
(4) Karki, S.; Friščić, T.; Fábián, L.; Laity, P. R.; Day, G. M.; Jones, W., Improving Mechanical Properties of Crystalline Solids by Cocrystal Formation: New Compressible Forms of Paracetamol. *Adv. Mater.* **2009,** 21, 3905-3909.
(5) Gajda, M.; Nartowski, K. P.; Pluta, J.; Karolewicz, B., Continuous, one-step synthesis of pharmaceutical cocrystals via hot melt extrusion from neat to matrix-assisted processing – State of the art. *Int. J. Pharm.* **2019,** 558, 426-440.
(6) Barmpalexis, P.; Karagianni, A.; Nikolakakis, I.; Kachrimanis, K., Artificial neural networks (ANNs) and partial least squares (PLS) regression in the quantitative analysis of cocrystal formulations by Raman and ATR-FTIR spectroscopy. *J. Pharm. Biomed. Anal.* **2018,** 158, 214-224.
(7) Gajda, M.; Nartowski, K. P.; Pluta, J.; Karolewicz, B., The role of the polymer matrix in solvent-free hot melt extrusion continuous process for mechanochemical synthesis of pharmaceutical cocrystal. *Eur. J. Pharm. Biopharm.* **2018,** 131, 48-59.
(8) Dhumal, R. S.; Kelly, A. L.; York, P.; Coates, P. D.; Paradkar, A., Cocrystalization and Simultaneous Agglomeration Using Hot Melt Extrusion. *Pharm. Res.* **2010,** 27, 2725-2733.
(9) Boksa, K.; Otte, A.; Pinal, R., Matrix-Assisted Cocrystallization (MAC) Simultaneous Production and Formulation of Pharmaceutical Cocrystals by Hot-Melt Extrusion. *J. Pharm. Sci.* **2014,** 103, 2904-2910.
(10) Korde, S.; Pagire, S.; Pan, H.; Seaton, C.; Kelly, A.; Chen, Y.; Wang, Q.; Coates, P.; Paradkar, A., Continuous Manufacturing of Cocrystals Using Solid State Shear Milling Technology. *Cryst. Growth Des.* **2018,** 18, 2297-2304.
(11) Barmpalexis, P.; Koutsidis, I.; Karavas, E.; Louka, D.; Papadimitriou, S. A.; Bikiaris, D. N., Development of PVP/PEG mixtures as appropriate carriers for the preparation of drug solid dispersions by melt mixing technique and optimization of dissolution using artificial neural networks. *Eur. J. Pharm. Biopharm.* **2013,** 85, 1219-1231.
(12) Hasa, D.; Jones, W., Screening for new pharmaceutical solid forms using mechanochemistry: A practical guide. *Advanced Drug Delivery Reviews* **2017,** 117, 147-161.
(13) Douroumis, D.; Ross, S. A.; Nokhodchi, A., Advanced methodologies for cocrystal synthesis. *Adv. Drug Delivery Rev.* **2017,** 117, 178-195.

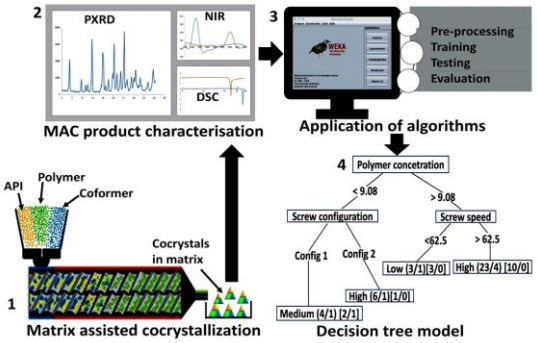
(14) Grymonpré, W.; Bostijn, N.; Herck, S. V.; Verstraete, G.; Vanhoorne, V.; Nuhn, L.; Rombouts, P.; Beer, T. D.; Remon, J. P.; Vervaet, C., Downstream processing from hot-melt extrusion towards tablets: A quality by design approach. *Int. J. Pharm.* **2017,** 531, 235-245.
(15) Witten, I. H.; Frank, E.; Hall, M. A.; Pal, C. J., *Data mining: practical machine learning tools and techniques*. Fourth / Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal.;Fourth; ed.; Morgan Kaufmann: Amsterdam, 2017.
(16) Kalmegh, S., Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news. *Int. j. innov. res. sci. eng.* **2015,** 2, 438-446.
(17) Mistry, P.; Neagu, D.; Trundle, P. R.; Vessey, J. D., Using random forest and decision tree models for a new vehicle prediction approach in computational toxicology. *Soft Comput.* **2016,** 20, 2967-2979.
(18) Ronowicz, J.; Thommes, M.; Kleinebudde, P.; Krysiński, J., A data mining approach to optimize pellets manufacturing process based on a decision tree algorithm. *Eur. J. Pharm. Sci.* **2015,** 73, 44-48.
(19) Rebecca Jeya Vadhanam, B.; Mohan, S.; Ramalingam, V. V.; Sugumaran, V., Performance Comparison of Various Decision Tree Algorithms for Classification of Advertisement and Non Advertisement Videos. *Indian. J. Sci. Technol.* **2016,** 9, 1-12.
(20) Snousy, M. B. A.; El-Deeb, H. M.; Badran, K.; Khlil, I. A. A., Suite of decision tree-based classification algorithms on cancer gene expression data. *Egypt. Inform. J.* **2011,** 12, 73-82.
(21) Mishra, A. K.; Ratha, B. K., Study of random tree and random forest data mining algorithms for microarray data analysis. *Int. J. Adv. Elect. Comput. Eng.* **2016,** 3, 5-7.
(22) Ali, J.; Khan, R.; Ahmad, N.; Maqsood, I., Random forests and decision trees. *Int. J. Comput. Integ. M.* **2012,** 9, 272.
(23) Djuris, J., *Computer-aided applications in pharmaceutical technology*. ed.; Elsevier: 2013.
(24) Neagu, D.; Richarz, A. N., *Big data in predictive toxicology*. ed.; Royal Society of Chemistry: 2019.
(25) Ismail, H. Y.; Singh, M.; Darwish, S.; Kuhs, M.; Shirazian, S.; Croker, D. M.; Khraisheh, M.; Albadarin, A. B.; Walker, G. M., Developing ANN-Kriging hybrid model based on process parameters for prediction of mean residence time distribution in twin-screw wet granulation. *Powder Technol.* **2019,** 343, 568-577.
(26) Sajjia, M.; Shirazian, S.; Kelly, C. B.; Albadarin, A. B.; Walker, G., ANN Analysis of a Roller Compaction Process in the Pharmaceutical Industry. *Chem. Eng. Technol.* **2017,** 40, 487-492.
(27) Maheshwari, S.; Agrawal, J.; Sharma, S., New approach for classification of highly imbalanced datasets using evolutionary algorithms. *Int. J. Sci. Eng. Res* **2011,** 2, 1-5.
(28) Analytics vidhya How to handle Imbalanced Classification Problems in machine learning? https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/ (14 September 2019),
(29) Narkhede, S. Understanding AUC - ROC Curve. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (19 September 2019),
(30) Narkhede, S. Understanding Confusion Matrix. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 (19 September 2019),
(31) Brownlee, J. What is a Confusion Matrix in Machine Learning. https://machinelearningmastery.com/confusion-matrix-machine-learning/ (19 September 2019),

(32) Crowley, M. M.; Zhang, F.; Koleng, J. J.; McGinity, J. W., Stability of polyethylene oxide in matrix tablets prepared by hot-melt extrusion. *Biomaterials* **2002,** 23, 4241-4248.
(33) Ibrahim, A. Y.; Forbes, R. T.; Blagden, N., Spontaneous crystal growth of co-crystals: the contribution of particle size reduction and convection mixing of the co-formers. *CrystEngComm* **2011,** 13, 1141-1152.
(34) Kelly, A. L.; Gough, T.; Dhumal, R. S.; Halsey, S. A.; Paradkar, A., Monitoring ibuprofen–nicotinamide cocrystal formation during solvent free continuous cocrystallization (SFCC) using near infrared spectroscopy as a PAT tool. *Int. J. Pharm.* **2012,** 426, 15-20.

“For Table of Contents Only”

Understanding Matrix Assisted Continuous Cocrystallisation using Data Mining approach in Quality by Design (QbD)

Billy Chabalenge[2], Sachin Korde[1,2], Adrian L Kelly[1,3], Daniel Neagu[3], Anant Paradkar[1,2]*



Synopsis

It is demonstrated that Artificial Intelligence based QbD approach can help in better understanding and decision making for effective matrix assisted cocrystallisation using hot melt extrusion.